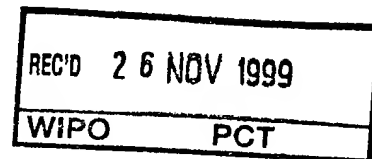


DE 99/2846



**PRIORITY DOCUMENT**  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH  
RULE 17.1(a) OR (b)

**Bescheinigung**

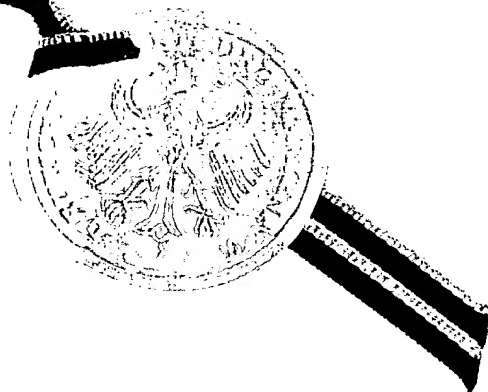
Die Siemens Aktiengesellschaft in München/Deutschland hat eine Patentanmeldung  
unter der Bezeichnung

"Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen  
für ein System, welches Zustände aufweist, wobei ein Zustands-  
übergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt"

am 23. September 1998 beim Deutschen Patent- und Markenamt eingereicht.

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprüng-  
lichen Unterlagen dieser Patentanmeldung.

Die Anmeldung hat im Deutschen Patent- und Markenamt vorläufig das Symbol  
G 05 B 17/00 der Internationalen Patentklassifikation erhalten.



München, den 21. Oktober 1999

**Deutsches Patent- und Markenamt**

**Der Präsident**

Im Auftrag

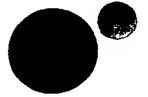
Jerofsky

Aktenzeichen: 198 43 620.3



.

.





## Beschreibung

Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt

Die Erfindung betrifft ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt.

Ein solches Verfahren und eine solche Anordnung sind aus [1] bekannt.

In [1] ist als Beispiel für ein solches System, welches Zustände aufweist, ein Finanzmarkt beschrieben.

Das System wird als ein Markov-Entscheidungsproblem beschrieben (Markov-Decision-Problem, MDP). Ein System, welches als Markov-Entscheidungsproblem beschrieben werden kann, ist in seiner Struktur in Fig.2 dargestellt.

Zu einem Zeitpunkt  $t$  befindet sich das System 201 in einem Zustand  $x_t$ . Der Zustand  $x_t$  ist für einen Beobachter des Systems beobachtbar. Aufgrund einer Aktion  $a_t$  aus einer Menge in dem Zustand  $x_t$  möglicher Aktionen,  $a_t \in A(x_t)$  geht das System mit einer gewissen Wahrscheinlichkeit in einen Folgezustand  $x_{t+1}$  zu einem Folgezeitpunkt  $t+1$  über.

Dies ist durch eine Schleife in Fig.2 symbolisch dargestellt. Ein Beobachter 200 nimmt beobachtbare Größen über den Zustand  $x_t$  wahr 202 und trifft eine Entscheidung über eine Aktion 203, mit der er auf das System 201 einwirkt. Das System 201 unterliegt üblicherweise einer Störung 205.

Ferner erhält der Beobachter 200 einen Gewinn  $r_t$  204

$$r_t = r(x_t, a_t, x_{t+1}) \in \mathcal{R}, \quad (1)$$

5 der von der Aktion  $a_t$  203 und dem ursprünglichen Zustand  $x_t$   
zu dem Zeitpunkt  $t$  sowie dem Folgezustand  $x_{t+1}$  des Systems zu  
dem Folgezeitpunkt  $t+1$  abhängt.

10 Der Gewinn  $r_t$  kann einen positiven oder negativen skalaren  
Wert annehmen, je nachdem, ob die Entscheidung zu einer hin-  
sichtlich eines vorgebbaren Kriteriums positiven oder negati-  
ven Systementwicklung führt, in [1] zu einer Kapitalvermehr-  
ung oder zu einem Verlust.

15 In einem weiteren Zeitschritt entscheidet sich der Beobachter  
200 des Systems 201 aufgrund der beobachtbaren Größen 202,  
204 des Folgezustandes  $x_{t+1}$  für eine neue Aktion  $a_{t+1}$  usw.

Eine Folge von

Zustand:	$x_t$	$\in$	$X$
Aktion:	$a_t$	$\in$	$A(x_t)$
Folgezustand:	$x_{t+1}$	$\in$	$X$
Gewinn	$r_t = r(x_t, a_t, x_{t+1})$	$\in$	$\mathcal{R}$

20

usw. beschreibt eine Trajektorie des Systems, die durch ein  
Performanzkriterium, das die einzelnen Gewinne  $r_t$  über die  
Zeitpunkte  $t$  akkumuliert, bewertet wird. Bei einem Markov-  
Entscheidungsproblem wird vereinfachend angenommen, daß der  
25 Zustand  $x_t$  und die Aktion  $a_t$  alle Informationen enthalten, um  
eine Übergangswahrscheinlichkeit  $p(x_{t+1}| \cdot)$  des Systems von dem  
Zustand  $x_t$  zu dem Folgezustand  $x_{t+1}$  zu beschreiben.

Formal bedeutet dies:

30

$$p(x_{t+1}|x_t, K, x_0, a_t, K, a_0) = p(x_{t+1}|x_t, a_t). \quad (2)$$

Mit  $p(x_{t+1}|x_t, a_t)$  wird eine Übergangswahrscheinlichkeit für den Folgezustand  $x_{t+1}$  bei gegebenem Zustand  $x_t$  und gegebener Aktion  $a_t$  bezeichnet.

5 Bei einem Markov-Entscheidungsproblem hängen also zukünftige  
Zustände des Systems nicht von Zuständen und Aktionen ab,  
die weiter als einen Zeitschritt in der Vergangenheit liegen.

Zusammenfassend sind im weiteren die Charakteristika eines  
10 Markov-Entscheidungsproblems dargestellt:

$X$	Menge der möglichen Zustände des Systems, z.B. $X = \mathcal{R}^m$ ,
$A(x_t)$	Menge der möglichen Aktionen in dem Zustand
$p(x_{t+1} x_t, a_t)$	$x_t$
$r(x_t, a_t, x_{t+1})$	Gewinn mit Erwartungswert $R(x_t, a_t)$ .

Das Ziel ist es, ausgehend von beobachtbaren Größen, den im  
weiteren als Trainingsdaten bezeichneten Größen, eine Strategie  
15 zu ermitteln, d.h. eine Folge von Funktionen

$$\pi = \{\mu_0, \mu_1, K, \mu_T\}, \quad (3)$$

welche zu jedem Zeitpunkt  $t$  jeden Zustand in eine Handlungs-  
20 vorschrift, d.h. Aktion

$$\mu_t(x_t) = a_t \quad (4)$$

abbilden.

25

Eine solche Strategie wird durch eine Optimierungsfunktion bewertet.

Die Optimierungsfunktion gibt den Erwartungswert, der über die Zeit akkumulierten Gewinne bei einer gegebenen Strategie  $\pi$  und einem Startzustand  $x_0$  an.

- 5 Als ein Beispiel eines Verfahrens des approximativen dynamischen Programmierens ist in [1] das sogenannte Q-Lernverfahren beschrieben.

Eine optimale Bewertungsfunktion  $V^*(x)$  ist definiert durch

10

$$V^*(x) = \max_{\pi} V^{\pi}(x) \quad \forall x \in X \quad (5)$$

mit

$$15 \quad V^{\pi}(x) = E \left[ \sum_{t=0}^{\infty} \gamma^t r(x_t, \mu_t, x_{t+1}) \mid x_0 = x \right], \quad (6)$$

wobei mit  $\gamma$  ein vorgegebbarer Verringerungsfaktor bezeichnet wird, der gemäß folgender Vorschrift gebildet wird:

$$20 \quad \gamma = \frac{1}{1+z}, \quad (7)$$

$$z \in \mathbb{R}^+. \quad (8)$$

- Im Rahmen des Q-Lernverfahrens wird für jedes Paar  
 25 (Zustand  $x_t$ , Aktion  $a_t$ ) eine Q-Bewertungsfunktion  $Q^*(x_t, a_t)$  gemäß folgender Vorschrift gebildet:

$$Q^*(x_t, a_t) = \sum_{x \in X} p(x_{t+1} | x_t, a_t) \cdot r_t + \gamma \cdot \sum_{x \in X} p(x | x_t, a_t) \cdot \max_{a \in A} (Q^*(x, a))$$

(9)

Aufgrund jeweils des Tupels  $(x_t, x_{t+1}, a_t, r_t)$  werden die  $Q$ -Werte  $Q^*(x, a)$  in der  $k+1$  ten Iteration gemäß folgender Lernregel mit einer vorgegebenen Lernrate  $\eta_k$  gemäß folgender Vorschrift adaptiert:

5

$$Q_{k+1}(x_t, a_t) = (1 - \eta_k)Q_k(x_t, a_t) + \eta_k \left( r_t + \gamma \max_{a \in A} (Q_k(x_{t+1}, a)) \right). \quad (10)$$

10 Üblicherweise werden die sogenannten  $Q$ -Werte  $Q^*(x, a)$  durch jeweils einen Funktionsapproximator, beispielsweise ein neuronales Netz oder auch einen Polynomklassifikator, mit einem Gewichtsvektor  $w^a$ , der Gewichte des Funktionsapproximators enthält, für verschiedene Aktionen  $a$  approximiert.

15 Unter einem Funktionsapproximator ist beispielsweise ein neuronales Netz, ein Polynomklassifikator oder auch eine Kombination eines neuronalen Netzes mit einem Polynomklassifikator zu verstehen.

20 Es gilt also:

$$Q^*(x, a) \approx Q(x; w^a). \quad (11)$$

25 Änderungen der Gewichte in dem Gewichtsvektor  $w^a$  basieren auf einer temporären Differenz  $d_t$ , die gemäß folgender Vorschrift gebildet wird:

$$d_t := r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}; w_k^a) - Q(x_t; w_k^{a_t}) \quad (12)$$

30 Es ergibt sich für das  $Q$ -Lernverfahren unter Verwendung eines neuronalen Netzes folgende Adaptionsvorschrift für die Gewichte des neuronalen Netzes, welche Gewichte in dem Gewichtsvektor  $w^a$  enthalten sind:

$$w_{k+1}^{a_t} = w_k^{a_t} + \eta_k \cdot d_t \cdot \nabla Q(x_t; w_k^{a_t}). \quad (13)$$

Unter Verwendung der Trainingsdaten, die als Zeitreihenwerte Informationen über vorangegangene Kursverläufe eines Finanz-

5 marktes beschreiben, wird das neuronale Netz, welches das System Finanzmarkt, wie in [1] beschrieben, darstellt, trainiert.

Ein weiteres Verfahren der approximativen dynamischen Programmierung, das sogenannte TD( $\lambda$ )-Lernverfahren, ist aus [2] 10 bekannt und wird im Zusammenhang mit einem Ausführungsbeispiel näher erläutert.

Ferner ist aus [3] bekannt, welches Risiko mit einer Strategie  $\pi$  und einem Ausgangszustand  $x_t$  verbunden ist. Ein Verfahren 15 zur Risikovermeidung ist ebenfalls aus [3] bekannt.

Bei dem aus [3] bekannten Verfahren wird folgende Optimierungsfunktion, welche auch als erweiterte Q-Funktion 20  $\underline{Q}^\pi(x_t, a_t)$  bezeichnet wird, verwendet:

maximiere

$$\left( \underline{Q}^\pi(x_t, a_t) := r(x_t, a_t, x_{t+1}) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{ \sum_{k=1}^{\infty} \gamma^k r(x_k, \pi(x_k), x_{k+1}) \right\} \right)$$

(14)

25

Die erweiterte Q-Funktion  $\underline{Q}^\pi(x_t, a_t)$  beschreibt den schlechtesten Fall, falls in dem Zustand  $x_t$  die Aktion  $a_t$  ausgeführt wird und die Strategie  $\pi$  daraufhin verfolgt wird.

30 Die Optimierungsfunktion  $\underline{Q}^\pi(x_t, a_t)$  für



$$\underline{Q}^*(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^\pi(x_t, a_t) \quad (15)$$

ist gegeben, durch folgende Vorschrift:

$$\underline{Q}^*(x_t, a_t) = \min_{\substack{x \in X \\ p(x_{t+1}|x_t, a_t) > 0}} \left( r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^*(x, a) \right). \quad (16)$$

Ein erheblicher Nachteil dieser Vorgehensweise ist darin zu sehen, daß nur der schlechteste Fall im Rahmen der Strategiefindung berücksichtigt wird. Dies spiegelt jedoch die Anforderungen verschiedenster technischer Systeme nur in unzureichendem Ausmaß wieder.

15 Aus [4] ist es ferner bekannt, eine Zugangskontrolle für ein Kommunikationsnetz sowie das Routing innerhalb des Kommunikationsnetzes als ein Problem der dynamischen Programmierung zu formulieren.

20 Somit liegt der Erfindung das Problem zugrunde, ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System anzugeben, bei dem bzw. bei der eine erhöhte Flexibilität bei der Ermittlung der Strategie erreicht wird.

25 Das Problem wird durch das Verfahren sowie durch die Anordnung gemäß den Merkmalen der unabhängigen Patentansprüche gelöst.

30 Bei einem Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, erfolgt die Ermittlung der Folge von Aktionen derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungs-

funktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

5

Eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, weist einen Prozessor auf, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

Durch die Erfindung wird es erstmals möglich, in frei vorgebarbarer Genauigkeit im Rahmen einer Strategiefindung für eine möglichen Regelung oder Steuerung, allgemein einer Beeinflussung des Systems, ein Verfahren zur Ermittlung einer Folge von Aktionen anzugeben.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die im weiteren beschriebenen Weiterbildungen gelten sowohl für das Verfahren als auch die Anordnung, wobei bei der Weiterbildung der Anordnung jeweils der Prozessor derart eingerichtet ist, daß die Weiterbildung realisierbar ist.

In einer bevorzugten Ausgestaltung wird zur Ermittlung ein Verfahren des approximativen dynamischen Programmierens eingesetzt, beispielsweise ein auf dem Q-Lernen basierendes Verfahren oder auch ein auf dem TD( $\lambda$ )-Lernen basierendes Verfahren.

Im Rahmen des Q-Lernens wird bevorzugt die Optimierungsfunktion OFQ gemäß folgender Vorschrift gebildet:

$$\text{OFQ} = Q(x; w^a),$$

5

wobei mit

- $x$  ein Zustand in einem Zustandsraum  $X$ ,
- $a$  eine Aktion aus einem Aktionsraum  $A$ ,
- 10 •  $w^a$  die zur Aktion  $a$  gehörigen Gewichte eines Funktionsapproximators

bezeichnet wird/werden.

- 15 Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte  $w^a$  des Funktionsapproximators folgender Adaptionsschritt ausgeführt:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \delta_t \cdot \nabla Q(x_t; w_t^{a_t})$$

20

mit der Abkürzung

$$\delta_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^{a_t})$$

- 25 wobei mit

- $x_t, x_{t+1}$  jeweils ein Zustand in dem Zustandsraum  $X$ ,
- $a_t$  eine Aktion aus einem Aktionsraum  $A$ ,
- $\gamma$  ein vorgebbarer Verringerungsfaktor,
- 30 •  $w_t^{a_t}$  der zur Aktion  $a_t$  gehörige Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}^{a_t}$  der zur Aktion  $a_t$  gehörige Gewichtsvektor nach dem Adaptionsschritt,
- $\eta_t$  ( $t = 1, \dots$ ) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$  ein Risikokontrollparameter,
  - $\aleph^K$  eine Risikokontrollfunktion  $\aleph^K(\xi) = (1 - \kappa \text{sign}(\xi))\xi$ ,
  - $\nabla Q(\cdot)$  die Ableitung des Funktionsapproximators nach seinen Gewichten,
- 5 •  $r(x_t, a_t, x_{t+1})$  ein Gewinn beim Zustandsübergang von dem Zustand  $x_t$  nach dem Folgezustand  $x_{t+1}$ ,
- 

bezeichnet wird/werden.

- 10 Im Rahmen des  $TD(\lambda)$ -Lernverfahrens wird die Optimierungsfunktion bevorzugt gemäß folgender Vorschrift gebildet:

$$OFTD = J(x; w)$$

- 15 wobei mit

- $x$  ein Zustand in einem Zustandsraum  $X$ ,
- $a$  eine Aktion aus einem Aktionsraum  $A$ ,
- $w$  die Gewichte eines Funktionsapproximators

- 20 bezeichnet wird/werden.

- Im Rahmen des  $TD(\lambda)$ -Lernens wird zur Ermittlung der optimalen Gewichte  $w$  des Funktionsapproximators folgender Adaptions-
- 25 schritt ausgeführt:

$$w_{t+1} = w_t + \eta_t \cdot \aleph^K(d_t) \cdot z_t$$

mit den Abkürzungen

- 30  $d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

- 35  $z_{-1} = 0,$

wobei mit

- $x_t, x_{t+1}$  jeweils ein Zustand in dem Zustandsraum  $X$ ,
- $a_t$  eine Aktion aus einem Aktionsraum  $A$ ,
- 5 •  $\gamma$  ein vorgegebbarer Verringerungsfaktor,
- $w_t$  der Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}$  der Gewichtsvektor nach dem Adaptionsschritt,
- $\eta_t$  ( $t = 1, \dots$ ) eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$  ein Risikokontrollparameter,
- 10 •  $\aleph^K$  eine Risikokontrollfunktion  $\aleph^K(\xi) = (1 - \kappa \text{sign}(\xi))\xi$ ,
- $\nabla J(\cdot)$  die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$  ein Gewinn beim Zustandsübergang von dem Zustand  $x_t$  nach dem Folgezustand  $x_{t+1}$ ,

15

bezeichnet wird/werden.

Das System ist bevorzugt ein technisches System, von dem vor der Ermittlung Meßgrößen gemessen werden, die bei der Ermittlung der Folge von Aktionen verwendet werden.

20

Unter Verwendung der ermittelten Folge von Aktionen kann das technische System gesteuert oder auch geregelt werden.

Bevorzugt wird das System als ein Markov-Entscheidungsproblem modelliert.

Das Verfahren oder die Anordnung werden bevorzugt in einem Verkehrsleitsystem oder in einem Kommunikationssystem eingesetzt, wobei in dem Kommunikationssystem die Folge von Aktionen zur Durchführung einer Zugangskontrolle oder eines Routings, also einer Pfadvergabe, in einem Kommunikationsnetz eingesetzt wird.

30

Ferner kann das System ein Finanzmarkt sein, welcher durch ein Markov-Entscheidungsproblem modelliert wird und wobei der Verlauf des Finanzmarkts, beispielsweise ein Verlauf eines

35

Aktienindex oder auch ein Kursverlauf eines Devisenmarktes unter Verwendung der Verfahren bzw. der Anordnung analysiert und in den Markt entsprechend der Folge ermittelter Aktionen eingegriffen werden kann.

5

Ausführungsbeispiele der Erfindung sind in den Figuren dargestellt und werden im weiteren näher erläutert.

Es zeigen

10

Figur 1 ein Ablaufdiagramm, in dem einzelne Verfahrensschritte des ersten Ausführungsbeispiels dargestellt sind;

15

Figur 2 eine Skizze eines Systems, welches als Markov-Entscheidungsproblem modelliert werden kann;

20

Figur 3 eine Skizze eines Kommunikationsnetzes, bei dem in einer Vermittlungseinheit eine Zugangskontrolle durchgeführt wird;

Figur 4 eine symbolische Skizze eines Funktionsapproximators, mit dem ein Verfahren des approximativen dynamischen Programmierens realisiert wird;

25

Figur 5 eine weitere Skizze von mehreren Funktionsapproximatoren, mit dem ein approximatives dynamisches Programmieren implementiert wird;

30

Figur 6 eine Skizze eines Verkehrsleitsystems, welches gemäß einem Ausführungsbeispiel geregelt wird.

**Erstes Ausführungsbeispiel: Zugangskontrolle und Routing.**

**Fig.3** zeigt ein Kommunikationsnetz 300, welches eine Vielzahl von Vermittlungseinheiten 301a, 301b, ..., 301i, ... 301n aufweist, die über Verbindungen 302a, 302b, 302j, ... 302m miteinander verbunden sind.

---

10 Ferner ist ein erstes Endgerät 303 mit einer ersten Vermittlungseinheit 301a verbunden. Von dem ersten Endgerät 303 wird eine Anforderungsnachricht 304 an die erste Vermittlungseinheit 301a gesendet, mit der eine Reservierung einer vorgegebenen Bandbreite innerhalb des Kommunikationsnetzes 300 zur Übertragung von Daten (Videodaten, textuelle Daten) angefordert wird.

15

In der ersten Vermittlungseinheit 301a wird gemäß einer im weiteren beschriebenen Strategie ermittelt, ob die angeforderte Bandbreite in dem Kommunikationsnetz 300 auf einer angegebenen, angeforderten Verbindung verfügbar ist  
20 (Schritt 305).

Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 306).

25 Ist ausreichend Bandbreite verfügbar, so wird in einem weiteren Überprüfungsschritt (Schritt 307) überprüft, ob die Bandbreite reserviert werden kann.

30 Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 308).

Sonst wird von der ersten Vermittlungseinheit 301a eine Route von der ersten Vermittlungseinheit 301a über weitere Vermittlungseinheiten 301i zu einem zweiten Endgerät 309, mit dem  
35 das erste Endgerät 303 kommunizieren will, ausgewählt und es wird eine Verbindung initialisiert (Schritt 310).

Im folgenden wird von einem Kommunikationsnetz 300 ausgegangen, welches einen Satz von Vermittlungseinheiten

$$N = \{1, K, n, K, N\} \quad (17)$$

5 und einen Satz von physikalischen Verbindungen

---

$$L = \{1, K, l, K, L\}, \quad (18)$$

umfaßt, wobei eine physikalische Verbindung  $l$  eine Kapazität  
10 von  $B(l)$  Bandbreiteneinheiten aufweist.

Es sind ein Satz

$$M = \{1, K, m, K, M\} \quad (19)$$

15

verschiedener Diensttypen  $m$  verfügbar, wobei ein Diensttyp  $m$  durch

- einen Bandbreitenbedarf  $b(m)$ ,
  - eine durchschnittliche Verbindungsdauer  $\frac{1}{v(m)}$ , und
  - 20 • einen Gewinn  $c(m)$ , den man dann erhält, wenn eine Verbindungsanforderung des entsprechenden Diensttyps  $m$  akzeptiert wird,
- charakterisiert ist.

25 Der Gewinn  $c(m)$  ist gegeben durch die Menge des Geldes, die ein Netzwerkbetreiber des Kommunikationsnetzes 300 einem Teilnehmer für eine Verbindung des Diensttyps in Rechnung stellt. Anschaulich spiegelt der Gewinn  $c(m)$  unterschiedliche, von dem Netzbetreiber vorgebbare Prioritäten wider, die  
30 er mit verschiedenen Diensten assoziiert.

Eine physikalischen Verbindung  $l$  kann gleichzeitig eine beliebige Kombination von Kommunikationsverbindungen bereitstellen, solange die genutzte Bandbreite der Kommunikations-  
35 verbindungen nicht die insgesamt verfügbare Bandbreite der physikalischen Verbindung übersteigt.



Wird eine neue Kommunikationsverbindung des Typs  $m$  angefordert zwischen einem ersten Knoten  $i$  und einem zweiten Knoten  $j$  (Endgeräte werden auch als Knoten bezeichnet), so kann die angeforderte Kommunikationsverbindung, wie oben dargestellt, entweder akzeptiert oder zurückgewiesen werden.

Wird die Kommunikationsverbindung akzeptiert, so wird eine Route aus einer Menge vorgegebener Routen ausgewählt. Diese Auswahl wird als Routing bezeichnet. Im Rahmen der Kommunikationsverbindung vom Typ  $m$  werden  $b(m)$  Bandbreiteneinheiten für jede physikalische Verbindung entlang der ausgewählten Route für die Verbindungsdauer benutzt.

Somit kann im Rahmen der Zugangskontrolle (Call-Admission-Control) eine Route innerhalb des Kommunikationsnetzes 300 nur ausgewählt werden, wenn die ausgewählte Route ausreichend Bandbreite zur Verfügung hat.

Ziel der Zugangskontrolle und des Routings ist es, einen langfristigen Gewinn, der durch Akzeptanz der angeforderten Verbindungen erhalten wird, zu maximieren.

Das technische System Kommunikationsnetz 300 befindet sich zu einem Zeitpunkt  $t$  in einem Zustand  $x_t$ , welcher durch eine Liste von Routen über bestehende Verbindungen beschrieben wird, durch welche Listen angezeigt ist, wie viele Verbindungen welchen Diensttyps zu dem Zeitpunkt  $t$  die jeweilige Route verwenden.

Ereignisse  $w$ , durch die ein Zustand  $x_t$  in einen Folgezustand  $x_{t+1}$  überführt werden könnte, sind das Ankommen neuer Verbindungsanforderungsnachrichten oder auch das Beenden einer in dem Kommunikationsnetz 300 bestehenden Verbindung.

Eine Aktion  $a_t$  zu einem Zeitpunkt  $t$  aufgrund einer Verbindungsanforderung ist in diesem Ausführungsbeispiel die Ent-

scheidung, ob eine Verbindungsanforderung akzeptiert oder zurückgewiesen werden soll und, falls die Verbindung akzeptiert wird, die Auswahl der Route durch das Kommunikationsnetz 300.

- 5 Ziel ist die Ermittlung einer Folge von Aktionen, d.h. anschaulich das Lernen einer Strategie mit Aktionen zu einem Zustand  $x_t$  zu bestimmen derart, daß folgende Vorschrift maximiert wird:

$$10 \quad E \left( \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}) \right), \quad (20)$$

wobei mit

- $E\{. \}$  ein Erwartungswert,
- 15 •  $t_k$  ein Zeitpunkt, zu dem ein  $k$ -tes Ereignis erfolgt,
- $g(x_{t_k}, \omega_k, a_{t_k})$  der Gewinn, der mit dem  $k$ -ten Ereignis verbunden ist, und
- $\beta$  ein Verringerungsfaktor, der einen sofortigen Gewinn wertvoller bewertet als ein Gewinn in ferner in der Zukunft
- 20 liegenden Zeitpunkten,

bezeichnet wird.

- 25 Unterschiedliche Realisierungen einer Strategie führen üblicherweise zu unterschiedlichen Gesamtgewinnen  $G$ :

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}). \quad (21)$$

- 30 Ziel ist die Maximierung des Erwartungswerts des Gesamtgewinns  $G$  gemäß folgender Vorschrift  $J$ :

$$J = E \left\{ \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}) \right\}, \quad (22)$$

wobei ein Risiko, daß der Gesamtgewinn  $G$  einer speziellen Realisierung einer Zugangskontrolle und einer Routing-

5 Strategie unter den Erwartungswert sinkt, einstellbar ist.

Zur Durchführung der Zugriffskontrolle sowie zum Routing wird das  $TD(\lambda)$ -Lernverfahren eingesetzt.

10 Es wird folgende Zielfunktion im Rahmen dieses Ausführungsbeispiels eingesetzt:

$$J^*(x_t) = E_{\tau} \left\{ e^{-\beta \tau} \right\} E_{\omega} \left\{ \max_{a \in A} \left[ g(x_t, \omega_t, a) + J^*(x_{t+1}) \right] \right\}, \quad (23)$$

15 wobei mit

- $A$  ein Aktionsraum mit einer vorgegebenen Anzahl Aktionen, die in einem Zustand  $x_t$  jeweils zur Verfügung stehen,
- $\tau$  ein erster Zeitpunkt, zudem ein erstes Ereignis  $\omega$  erfolgt,
- $x_{t+1}$  ein Folgezustand des Systems,

20

bezeichnet wird.

25 Ein approximierter Wert des Zielwerts  $J^*(x_t)$  wird unter Verwendung eines Funktionsapproximators 400 (vgl. **Fig.4**) unter Verwendung von Trainingsdaten gelernt und gespeichert.

Trainingsdaten sind zuvor in dem Kommunikationsnetz 300 gemessene Daten über das Verhalten des Kommunikationsnetzes 300 bei ankommenden Verbindungsanforderungen 304 sowie bei Beendigung von Nachrichten. Diese zeitliche Folge von Zuständen wird gespeichert und mit diesen Trainingsdaten wird der Funktionsapproximator 400 gemäß dem im weiteren beschriebenen  
35 Lernverfahren trainiert.

Als Eingangsgröße des Funktionsapproximators 400 dienen für jeden Eingang 401, 402, 403 des Funktionsapproximators 400 jeweils eine Anzahl von Verbindungen jeweils eines Diensttyps m auf einer Route des Kommunikationsnetzes 300. Diese sind in **Fig. 4** durch Blöcke 404, 405, 406 symbolisch dargestellt.

Ausgangsgröße des Funktionsapproximators 400 ist ein approximierter Zielwert  $\tilde{J}$  des Zielwerts  $J^*$ .

10

Eine detailliertere Darstellung des Funktionsapproximators 500, welcher in diesem Fall mehrere Teil-Funktionsapproximatoren 510, 520 des Funktionsapproximators 500 aufweist, zeigt **Fig. 5**. Eine Ausgangsgröße ist der approximierter Zielwert  $\tilde{J}$ , der gemäß folgender Vorschrift gebildet wird:

15

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^L \tilde{J}^{(l)}(x_t^{(l)}, \Theta_t^{(l)}). \quad (24)$$

20

Die Eingangsgrößen der Teilfunktionsapproximatoren 510, 520, die an Eingängen 511, 512, 513 des ersten Teil-Funktionsapproximators 510 bzw. an Eingängen 521, 522 und 523 des zweiten Teilfunktionsapproximators 520 anliegen, sind wiederum jeweils eine Anzahl von Diensttypen eines Typs m jeweils in einer physikalischen Verbindung r, symbolisiert durch Blöcke 514, 515, 516 für den ersten Teil-Funktionsapproximator und 524, 525 und 526 für den zweiten Teil-Funktionsapproximator 520.

25

30

Teilausgangsgrößen 530, 531, 532, 533 werden einer Addiereinheit 540 zugeführt und als Ausgangsgröße der Addiereinheit wird die approximierter Zielgröße  $\tilde{J}$  gebildet.

35

Angenommen, das Kommunikationsnetz 300 befindet sich in dem Zustand  $x_{t_k}$  und eine Anforderungsnachricht, mit der ein Diensttyp m der Klasse m für eine Verbindung zwischen zwei

Knoten  $i, j$  angefordert wird, gelangt zu der ersten Verbindungseinheit 301a.

Mit  $R(i, j)$  wird eine Liste erlaubter Routen zwischen den  
5 Knoten  $i$  und  $j$  bezeichnet und mit

$$\tilde{R}(i, j, x_{t_k}) \subset R(i, j) \quad (25)$$

10 wird eine Liste aller möglichen Routen als Teilmenge der Routen  $R(i, j)$  bezeichnet, die hinsichtlich der verfügbaren und angeforderten Bandbreite eine mögliche Verbindung realisieren könnten.

Für jede mögliche Route  $r, r \in \tilde{R}(i, j, x_{t_k})$  wird ein Folgezustand  $x_{t_k+1}(x_{t_k}, \omega_k, r)$  ermittelt, der daraus resultiert, daß  
15 die Verbindungsanforderung 304 akzeptiert wird und die Verbindung auf der Route  $r$  dem anfordernden ersten Endgerät 303 zur Verfügung gestellt wird.

20 Dies ist in **Fig.1** als zweiter Schritt (Schritt 102) dargestellt, wobei in einem ersten Schritt (Schritt 101) jeweils der Zustand des Systems sowie das jeweilige Ereignis festgestellt werden.

25 Es wird in einem dritten Schritt (Schritt 103) eine auszuwählende Route  $r^*$  gemäß folgender Vorschrift ermittelt:

$$r^* = \arg \max_{r \in \tilde{R}(i, j, x_{t_k})} \tilde{J}(x_{t_k+1}(x_{t_k}, \omega_k, r), \Theta_t). \quad (26)$$

30 In einem weiteren Schritt (Schritt 104) wird überprüft, ob folgende Vorschrift erfüllt ist:

$$c(m) + \tilde{J}(x_{t_k+1}(x_{t_k}, \omega_k, r^*), \Theta_t) < \tilde{J}(x_{t_k}, \Theta_t). \quad (27)$$

Ist dies der Fall, so wird die Verbindungsanforderung 304 zurückgewiesen (Schritt 105), sonst wird die Verbindung akzeptiert und entlang der ausgewählten Route  $r^*$  zu dem Knoten  $j$  „durchgeschaltet“ (Schritt 106).

5

In einem Parametervektor  $\Theta$  sind jeweils für einen Zeitpunkt  $t$  Gewichte des Funktionsapproximators 400, 500 gespeichert, die im Rahmen des TD( $\lambda$ )-Lernverfahrens an die Trainingsdaten adaptiert werden, so daß eine optimierte Zugangskontrolle und  
10 ein optimiertes Routing erreicht wird.

Während der Trainingsphase werden die Gewichtsparameter an die dem Funktionsapproximator angelegten Trainingsdaten angepaßt.

15

Es wird ein Risikoparameter  $\kappa$  definiert, mit dem ein gewünschtes Risiko, welches durch eine Folge von Aktionen und Zuständen hinsichtlich eines vorgegebenen Zustands des Systems aufweist, einstellbar ist, gemäß folgenden Vorschriften:  
20

$-1 \leq \kappa < 0$ : risikoreiches Lernen,

$\kappa = 0$ : hinsichtlich des Risikos ein neutrales Lernen,

25

$0 < \kappa < 1$ : ein risiko-vermeidendes Lernen,

$\kappa = 1$ : „Worst-Case“-Lernen.

30

Ferner wird im Rahmen des Lernverfahrens ein vorgebbbarer Parameter  $0 \leq \lambda \leq 1$  und eine Schrittweitenfolge  $\gamma_k$  vorgegeben.

Die Gewichtswerte des Gewichtsvektors  $\Theta$  werden aufgrund jedes Ereignisses  $\omega_{t_k}$  gemäß folgender Adaptionvorschrift an

35 die Trainingsdaten angepaßt:

$$\Theta_k = \Theta_{k-1} + \gamma_k \delta_k \mathbf{z}_t, \quad (28)$$

wobei

$$d_k = e^{-\beta(t_k - t_{k-1})} \left( g(x_{t_k}, \omega_k, a_{t_k}) + \tilde{J}(x_{t_k}, \Theta_{k-1}) \right) - \tilde{J}(x_{t_{k-1}}, \Theta_{k-1})$$

5

(29)

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_{\Theta} \tilde{J}(x_{t_{k-1}}, \Theta_{k-1}), \quad (30)$$

und

10

$$\kappa^k(\xi) = (1 - \kappa \text{sign}(\xi)) \xi. \quad (31)$$

Es wird angenommen:  $z_{-1} = 0$ .

15 Die Funktion

$$g(x_{t_k}, \omega_k, a_{t_k}) \quad (32)$$

bezeichnet den sofortigen Gewinn gemäß folgender Vorschrift:

20

$$g(x_{t_k}, \omega_k, a_{t_k}) = \begin{cases} c(m) & \text{wenn } \omega_{t_k} \text{ ist eine Dienst an forderung eines} \\ & \text{Diensttyps } m \text{ und die Verbindung wird} \\ & \text{akzeptiert} \\ 0 & \text{sonst} \end{cases}$$

(33)

Es wird also, wie oben beschrieben, eine Folge von Aktionen  
 25 ermittelt, hinsichtlich einer Verbindungsanforderung, so daß  
 eine Verbindungsanforderung aufgrund einer Aktion entweder  
 zurückgewiesen oder akzeptiert wird. Die Ermittlung erfolgt  
 unter Berücksichtigung einer Optimierungsfunktion, in der das  
 Risiko mittels eines Risikokontrollparameters  $\kappa \in [-1; 1]$   
 30 variabel einstellbar ist.

**Zweites Ausführungsbeispiel: Verkehrsleitsystem**

**Fig. 6** zeigt eine Straße 600, die von Autos 601, 602, 603, 604, 605 und 606 befahren ist.

5

In die Straße 600 integrierte Leiterschleifen 610, 611 nehmen elektrische Signale in bekannter Weise auf und führen die elektrischen Signale 615, 616, einem Rechner 620 über eine Eingangs-/Ausgangsschnittstelle 621 zu. In einem mit der Eingangs-/Ausgangsschnittstelle 621 verbundenen Analog-/Digital-Wandler 622 werden die elektrischen Signale in eine Zeitreihe digitalisiert und in einem Speicher 623, der über einen Bus 624 mit dem Analog-/Digital-Wandler 622 und einem Prozessor 625 verbunden ist, gespeichert. Über die Eingangs-/Ausgangsschnittstelle 621 werden einem Verkehrsleitsystem 650 Steuerungssignale 651 zugeführt, aus denen in dem Verkehrsleitsystem 650 eine vorgegebene Geschwindigkeitsvorgabe 652 einstellbar ist oder auch weitere Angaben von Verkehrsvorschriften, die über das Verkehrsleitsystem 650 Fahrern der Fahrzeuge 601, 602, 603, 604, 605 und 606 dargestellt werden.

Zur Verkehrsmodellierung werden in diesem Fall folgende lokale Zustandsgrößen verwendet:

- Verkehrsflußgeschwindigkeit  $v$ ,
- 25 • Fahrzeugdichte  $\rho$  ( $\rho$  = Anzahl von Fahrzeugen pro Kilometer  $\frac{Fz}{km}$ ),
- Verkehrsfluß  $q$  ( $q$  = Anzahl der Fahrzeuge pro Stunde  $\frac{Fz}{h}$ ,  
( $q = v * \rho$ )), und
- jeweils zu einem Zeitpunkt von dem Verkehrsleitsystem 650  
30 angezeigte Geschwindigkeitsbegrenzungen 652.

Die lokalen Zustandsgrößen werden wie oben beschrieben unter Verwendung der Leiterschleifen 610, 611 gemessen.



Somit stellen diese Größen  $(v(t), p(t), q(t))$  einen Zustand des technischen Systems "Verkehr" zu einem bestimmten Zeitpunkt  $t$  dar.

- 5 In diesem Ausführungsbeispiel ist somit das System ein Verkehrssystem, welches unter Verwendung des Verkehrsleitsystems 650 geregelt wird.
- 

- 10 Als Verfahren des approximativen dynamischen Programmierens wird in diesem zweiten Ausführungsbeispiel ein erweitertes Q-Lernverfahren beschrieben.

Der Zustand  $x_t$  wird beschrieben durch einen Zustandsvektor

15 
$$x(t) = (v(t), p(t), q(t)). \quad (34)$$

Die Aktion  $a_t$  bezeichnet die Geschwindigkeitsbegrenzung 652, die zum Zeitpunkt  $t$  von dem Verkehrsleitsystem 650 angezeigt wird.

- 20 Der Gewinn  $r(x_t, a_t, x_{t+1})$  beschreibt die Güte des Verkehrsflusses, der zwischen den Zeitpunkten  $t$  und  $t+1$  von den Leierschleifen 610 und 611 gemessen wurde. Im Rahmen dieses zweiten Ausführungsbeispiels bezeichnet  $r(x_t, a_t, x_{t+1})$

- die mittlere Geschwindigkeit der Fahrzeuge im Zeitintervall  $[t, t + 1]$ ,

oder

- 30
- die Anzahl der Fahrzeuge, die im Zeitintervall  $[t, t + 1]$  die Leierschleifen 610 und 611 passiert haben,

oder

- 35
- die Varianz der Fahrzeuggeschwindigkeiten im Zeitintervall  $[t, t + 1]$ ,

oder

- eine gewichtete Summe aus den obigen Größen.

5

Für jede mögliche Aktion  $a_t$ , d.h. für jede von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung, wird ein Wert der Optimierungsfunktion OFQ ermittelt, wobei jeweils ein geschätzter Wert der Optimierungsfunktion OFQ als  
10 neuronales Netz realisiert wird.

15

Aus diese Weise ergibt sich eine Menge von Bewertungsgrößen für die unterschiedlichen Aktionen  $a_t$  in dem Systemzustand  $x_t$ .

In einer Regelungsphase wird aus den möglichen Aktionen  $a_t$ , d.h. aus der Menge der von dem Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen, diejenige Aktion  $a_t$  ausgewählt, für die in dem aktuellen Systemzustand  $x_t$  die maximale Bewertungsgröße OFQ ermittelt worden ist.  
20

Die aus dem Q-Lernverfahren bekannte Adaptionsvorschrift zur Berechnung der Optimierungsfunktion OFQ wird gemäß diesem Ausführungsbeispiel um eine Risikokontrollfunktion  $\kappa^k()$ , die  
25 das Risiko berücksichtigt, erweitert.

Wiederum wird der Risikokontrollparameter  $\kappa$  gemäß der Strategie aus dem ersten Ausführungsbeispiel im Intervall von  $[-1 \leq \kappa \leq 1]$  vorgegeben und repräsentiert das Risiko, das ein  
30 Benutzer im Rahmen der Anwendung hinsichtlich der zu bestimmenden Kontrollstrategie eingehen will.

Gemäß diesem Ausführungsbeispiel wird folgende Bewertungsfunktion OFQ verwendet:

35

$$\text{OFQ} = Q(x; w^a), \quad (35)$$

wobei mit

- $x = (v; p; q)$  ein Zustand des Verkehrssystems,
- 5 •  $a$  eine Geschwindigkeitsbegrenzung aus dem Aktionsraum  $A$  aller vom Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen,
- $w^a$  die zur Geschwindigkeitsbegrenzung  $a$  gehörigen Gewichte des neuronalen Netzes,

10

bezeichnet wird/werden.

Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte  $w^a$  den neuronalen Netzes folgender Adaptionsschritt  
 15 ausgeführt:

$$w_{t+1}^a = w_t^a + \eta_t \cdot \delta_t \cdot \nabla Q(x_t; w_t^a) \quad (36)$$

mit der Abkürzung

20

$$\delta_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^a) \quad (37)$$

wobei mit

25

- $x_t, x_{t+1}$  jeweils ein Zustand des Verkehrssystems gemäß Vorschrift (34),
- $a_t$  eine Aktion, d.h. eine von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung,
- $\gamma$  ein vorgegebbarer Verringerungsfaktor,
- 30 •  $w_t^a$  der zur Aktion  $a_t$  gehörige Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}^a$  der zur Aktion  $a_t$  gehörige Gewichtsvektor nach dem Adaptionsschritt,
- $\eta_t$  ( $t = 1, \dots$ ) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$  ein Risikokontrollparameter,
  - $N^K$  eine Risikokontrollfunktion  $N^K(\xi) = (1 - \kappa \text{sign}(\xi))\xi$ ,
  - $\nabla Q(\cdot)$  die Ableitung des neuronalen Netzes nach seinen Gewichten,
- 
- 5 •  $r(x_t, a_t, x_{t+1})$  ein Gewinn beim Zustandsübergang von dem Zustand  $x_t$  nach dem Folgezustand  $x_{t+1}$ ,

bezeichnet wird/werden.

- 10 Im Rahmen des Lernens kann aus den möglichen Aktionen  $a_t$  eine Aktion  $a_t$  zufällig gewählt werden. Hierbei ist es nicht erforderlich, die Aktion  $a_t$ , die zu der größten Bewertungsgröße geführt hat, zu wählen.
- 15 Die Adaption der Gewichte hat derart zu erfolgen, daß nicht nur eine im Erwartungswert der Optimierungsfunktion optimierte Regelung des Verkehrs erreicht wird, sondern auch eine Varianz der Regelungsergebnisse berücksichtigt wird.
- 20 Dies ist insbesondere vorteilhaft, da der Zustandsvektor  $x(t)$  das tatsächliche System Verkehr in einigen Aspekten nur unzureichend modelliert und es deshalb zu nicht erwarteten Störungen kommen kann. So hängt die Dynamik des Verkehrs und damit seiner Modellierung von weiteren Faktoren wie beispielsweise Wetter, Anteil an Lastkraftwagen auf der Straße, ein
- 25 Anteil von Wohnmobilen, etc. ab, die nicht immer in den Meßgrößen des Zustandsvektors  $x(t)$  integriert sind. Zudem ist nicht immer sichergestellt, daß die Verkehrsteilnehmer sofort den neuen Geschwindigkeitsangaben gemäß dem Verkehrsleitsystem Folge leisten.
- 30

Eine Regelungsphase an dem realen System gemäß dem Verkehrsleitsystem vollzieht sich gemäß folgenden Schritten:

- 35 1. Das Messen des Zustandes  $x_t$  zum Zeitpunkt  $t$  erfolgt an verschiedenen Stellen des Verkehrssystems Verkehr und ergibt einen Zustandsvektor  $x(t) = (v(t), \rho(t), q(t))$ .

2. Für alle möglichen Aktionen  $a_t$  wird ein Wert der Optimierungsfunktion ermittelt und es wird diejenige Aktion  $a_t$  mit der höchsten Bewertung in der Optimierungsfunktion ausgewählt.

---

In diesem Dokument sind folgende Veröffentlichungen zitiert:

- 5 [1] R. Neuneier, Enhancing Q-Learning for Optimal Asset Allocation, Proceedings of the Neural Information Processing Systems, NIPS 1997
- 
- [2] R.S. Sutton, Learning to predict by the method of temporal differences, Machine Learning, 3:9-44, 1988
- 10 [3] M. Heger, Risk and Reinforcement Learning: Concepts and Dynamic Programming, ZKW Bericht Nr. 8/94, Zentrum für Kognitionswissenschaften, Universität Bremen, ISSN 0947-0204, Dezember 1994
- 15 [4] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, Belmont, MA, 1995

**Patentansprüche**

1. Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein  
5 Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt,

bei dem die Ermittlung der Folge von Aktionen derart erfolgt, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

15 2. Verfahren nach Anspruch 1,  
bei dem zur Ermittlung ein Verfahren des approximativen Dynamischen Programmierens eingesetzt wird.

3. Verfahren nach Anspruch 2,  
20 bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem Q-Lernen basierendes Verfahren ist.

4. Verfahren nach Anspruch 3,  
bei dem die Optimierungsfunktion OFQ im Rahmen des Q-Lernens gemäß folgender Vorschrift gebildet wird:

$$\text{OFQ} = Q(x; w^a),$$

wobei mit

30

- $x$  ein Zustand in einem Zustandsraum  $X$ ,
- $a$  eine Aktion aus einem Aktionsraum  $A$ ,
- $w^a$  die zur Aktion  $a$  gehörigen Gewichte eines Funktionsapproximators

35

bezeichnet wird/werden, und bei dem die Gewichte des Funktionsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \kappa^K(d_t) \cdot \nabla Q(x_t; w_t^{a_t})$$

mit der Abkürzung

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^{a_t})$$

wobei mit

- 10 •  $x_t, x_{t+1}$  jeweils ein Zustand in dem Zustandsraum  $X$ ,
- $a_t$  eine Aktion aus einem Aktionsraum  $A$ ,
- $\gamma$  ein vorgebbarer Verringerungsfaktor,
- $w_t^{a_t}$  der zur Aktion  $a_t$  gehörige Gewichtsvektor vor dem Adaptionsschritt,
- 15 •  $w_{t+1}^{a_t}$  der zur Aktion  $a_t$  gehörige Gewichtsvektor nach dem Adaptionsschritt,
- $\eta_t$  ( $t = 1, \dots$ ) eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$  ein Risikokontrollparameter,
- $\kappa^K$  eine Risikokontrollfunktion  $\kappa^K(\xi) = (1 - \kappa \text{sign}(\xi))\xi$ ,
- 20 •  $\nabla Q(\cdot)$  die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$  ein Gewinn beim Zustandsübergang von dem Zustand  $x_t$  nach dem Folgezustand  $x_{t+1}$ ,

25 bezeichnet wird/werden.

5. Verfahren nach Anspruch 2,

bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem TD( $\lambda$ )-Lernen basierendes Verfahren ist.

30

6. Verfahren nach Anspruch 5,

bei dem die Optimierungsfunktion OFTD im Rahmen des TD( $\lambda$ )-Lernens gemäß folgender Vorschrift gebildet wird:



$$\text{OFTD} = J(x; w)$$

wobei mit

- 5
- $x$  ein Zustand in einem Zustandsraum  $X$ ,
  - $a$  eine Aktion aus einem Aktionsraum  $A$ ,
  - $w$  die Gewichte eines Funktionsapproximators
- 

10 bezeichnet wird/werden, und bei dem die Gewichte des Funktionsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1} = w_t + \eta_t \cdot \kappa^K(d_t) \cdot z_t$$

mit den Abkürzungen

15

$$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

20

$$z_{-1} = 0,$$

wobei mit

- $x_t, x_{t+1}$  jeweils ein Zustand in dem Zustandsraum  $X$ ,
  - $a_t$  eine Aktion aus einem Aktionsraum  $A$ ,
  - $\gamma$  ein vorgebbbarer Verringerungsfaktor,
  - $w_t$  der Gewichtsvektor vor dem Adaptionsschritt,
  - $w_{t+1}$  der Gewichtsvektor nach dem Adaptionsschritt,
  - $\eta_t$  ( $t = 1, \dots$ ) eine vorgebbare Schrittweitenfolge,
  - 30 •  $\kappa \in [-1; 1]$  ein Risikokontrollparameter,
  - $\kappa^K$  eine Risikokontrollfunktion  $\kappa^K(\xi) = (1 - \kappa \text{sign}(\xi))\xi$ ,
  - $\nabla J(\cdot)$  die Ableitung des Funktionsapproximators nach seinen Gewichten,
  - $r(x_t, a_t, x_{t+1})$  ein Gewinn beim Zustandsübergang von dem Zustand  $x_t$  nach dem Folgezustand  $x_{t+1}$ ,
- 35

bezeichnet wird/werden.

7. Verfahren nach einem der Ansprüche 1 bis 6,  
bei dem das System ein technisches System ist, von dem vor  
5 der Ermittlung Meßgrößen gemessen werden, die bei der Ermitt-  
lung der Folge von Aktionen verwendet werden.

---

8. Verfahren nach Anspruch 7,  
bei dem gemäß der Folge von Aktionen das technische System  
10 gesteuert wird.

9. Verfahren nach Anspruch 7,  
bei dem gemäß der Folge von Aktionen das technische System  
geregelt wird.

15 10. Verfahren nach einem der Ansprüche 1 bis 9,  
bei dem das System als ein Markov-Entscheidungsproblem model-  
liert wird.

20 11. Verfahren nach einem der Ansprüche 1 bis 10,  
eingesetzt in einem Verkehrsleitsystem.

12. Verfahren nach einem der Ansprüche 1 bis 10,  
eingesetzt in einem Kommunikationssystem.

25 13. Verfahren nach einem der Ansprüche 1 bis 10,  
eingesetzt zur Durchführung einer Zugangskontrolle in einem  
Kommunikationsnetz.

30 14. Verfahren nach einem der Ansprüche 1 bis 10,  
eingesetzt zur Durchführung eines Routings in einem Kommuni-  
kationsnetz.

15. Anordnung zur Ermittlung einer Folge von Aktionen für ein  
35 System, welches Zustände aufweist, wobei ein Zustandsübergang  
zwischen zwei Zuständen aufgrund einer Aktion erfolgt,

mit einem Prozessor, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

---

10 16. Anordnung nach Anspruch 15,  
eingesetzt zur Steuerung eines technischen Systems.

17. Anordnung nach Anspruch 15,  
eingesetzt zur Regelung eines technischen Systems.

15 18. Anordnung nach Anspruch 15,  
eingesetzt in einem Verkehrsleitsystem.

19. Anordnung nach Anspruch 15,  
20 eingesetzt in einem Kommunikationssystem.

20. Anordnung nach Anspruch 15,  
eingesetzt zur Durchführung einer Zugangskontrolle in einem Kommunikationsnetz.

21. Anordnung nach Anspruch 15,  
eingesetzt zur Durchführung eines Routings in einem Kommunikationsnetz.

### **Zusammenfassung**

**Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt**

---

Die Ermittlung der Folge von Aktionen erfolgt derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist. Die Optimierungsfunktion enthält einen variablen Parameter, mit dem ein Risiko einstellbar ist, welches Risiko die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist.

FIG 1

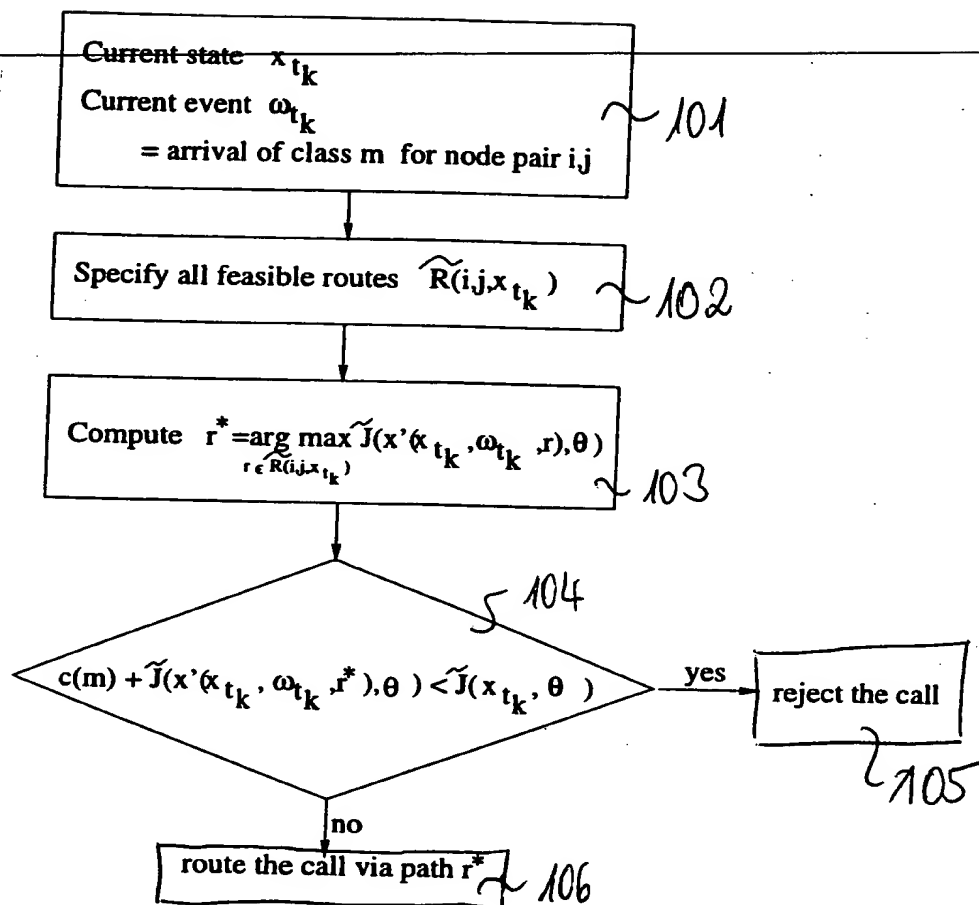


FIG 2

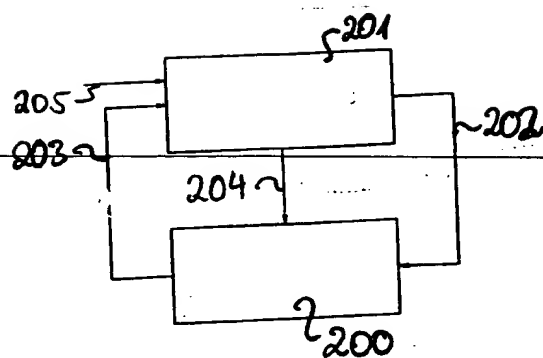


FIG 3

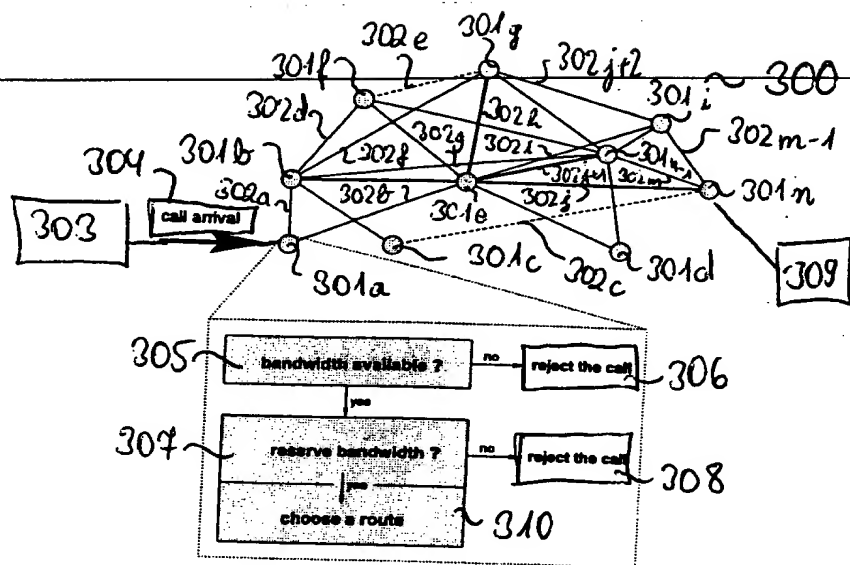


FIG 4

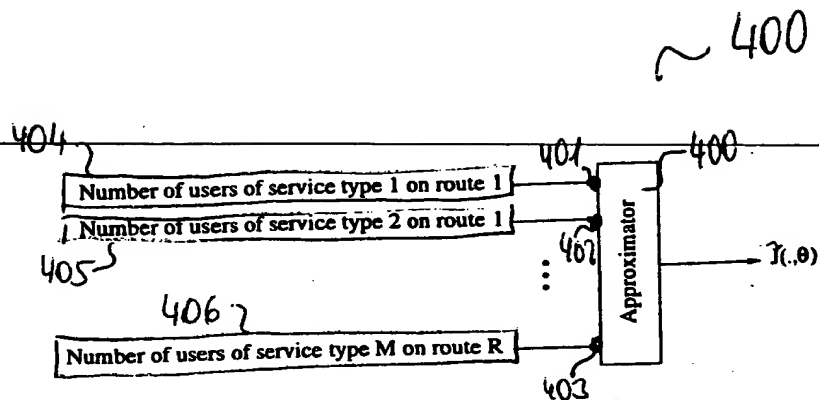


FIG 5

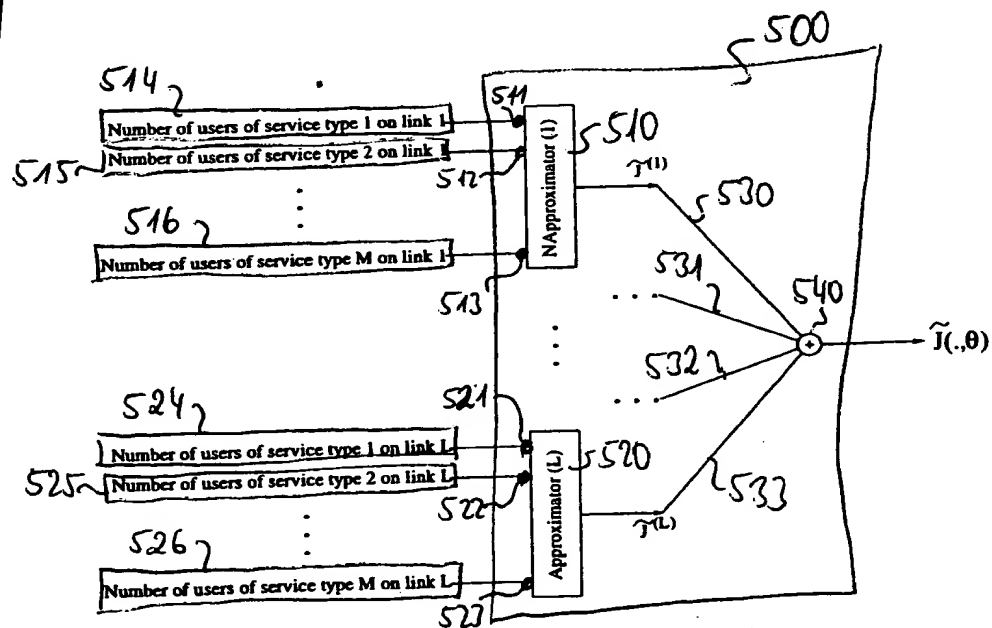
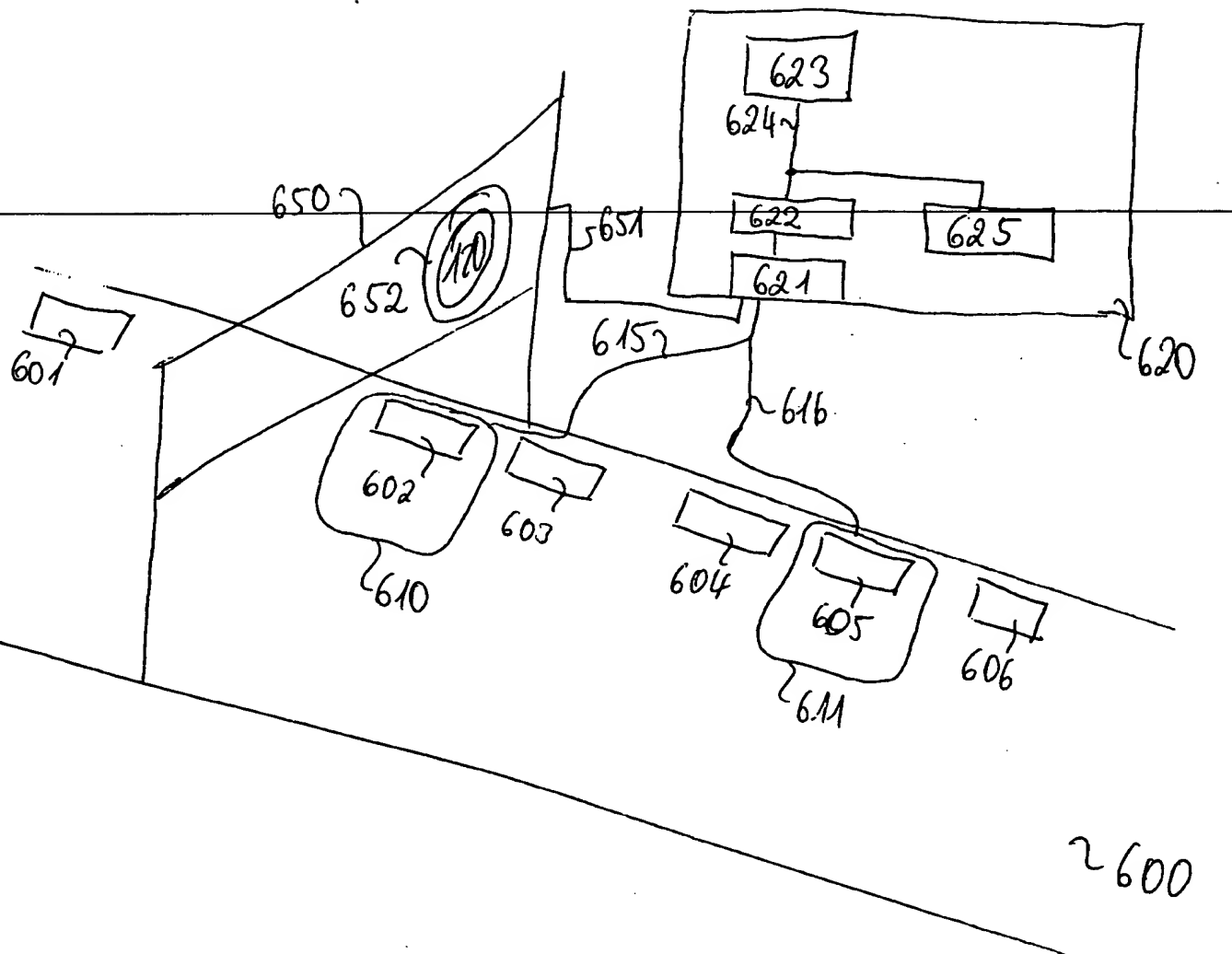




FIG 6



**THIS PAGE BLANK** (USPTO)

---